Differential Privacy Working Group Deliverables

Report of the CSACDifferential Privacy Working Group

Jay Breidt, CSAC Convener
Deborah Balk, John Czajka, Kathy Pettit, Allison Plyer (ex officio), Kunal Talwar, Richelle Winkler, Joe Whitley

September 18, 2020

Charter for CSAC Differential Privacy Working Group

Four tasks, with deliverables consisting of presentations at full CSAC meetings

- **Task 1.** Developing a summary of use cases.
- **Task 2.** Developing recommendations for prioritizing use cases for the administration of a "privacy-loss budget."
- **Task 3.** Developing metrics to assess the impact of differential privacy on the accuracy of decennial census data.
- **Task 4.** Developing strategies for communicating the use of differential privacy for the 2020 Census data products.

Outline of this presentation

- Timeline and process for this working group
- Brief background on differential privacy and Top Down Algorithm
- Findings on use cases for 2020 Census data products (Task 1)
- Findings on metrics for assessing fitness for use (Task 3)
- Findings on allocation of privacy-loss budget (Task 2)
- *Task 4 (communicating DP) is extremely important, but will be postponed to spring meeting, given the timeline*

Process and timeline

- WG approved on February 27, formed in the first week of March
  - NAC working group: same charter, same tasks, Dec. reporting
- Bureau held kickoff meeting plus "deep-dive" one-way briefings on Tasks 3 for CSAC/NAC working groups May 27- July 8
  - *WG would like to thank Bureau staff and subject-matter experts for briefings and follow-up Q&A*
- WG also met with one external subject-matter expert:
  - 08/13: Andrew Reamer, George Washington University
- WG has met regularly to finalize findings and draft recommendations:
  - 07/29, 08/17, 08/24, 08/31, 09/09, 09/14

Background

- From published 2010 census data, protected using “classic” Disclosure Avoidance Techniques (swapping, top and bottom coding, etc.), the Bureau…
  - **Reconstructed** block, sex, age, and ethnicity for 46% of the US population
  - **Re-identified** 38% of the reconstructed records by linking to commercial databases
  - (46%)(38%) = 17% of 2010 US population, or **52M people re-identified**
- *WG commends the Bureau for recognizing and demonstrating the vulnerability of classic Disclosure Avoidance Techniques: reconstruction/re-identification risks are serious and census data require protection*

Bureau’s alternative approach: Differential Privacy

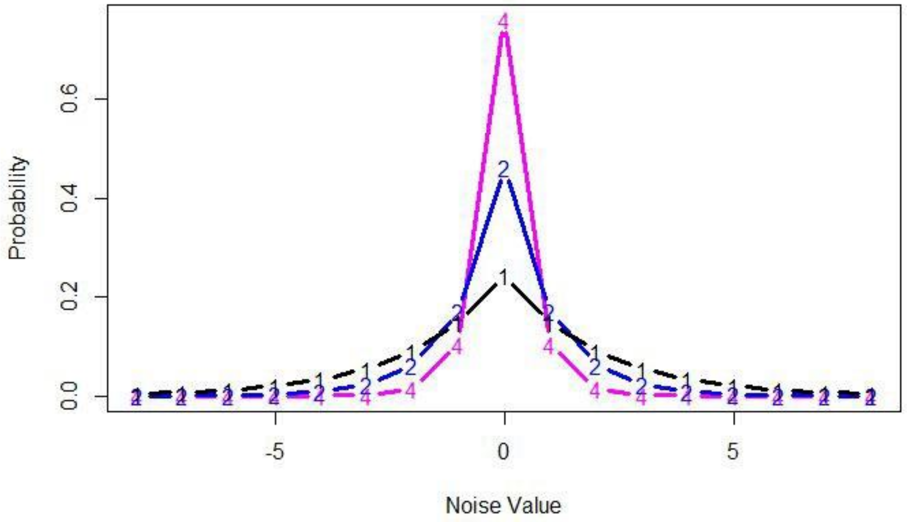
- Protecting against reconstruction attacks requires **limiting the number of queries** and **sacrificing some accuracy**
- DP adds noise to **any** published information from the original data:
  - Tabular summaries, microdata, metrics to assess fitness-for-use, ...
- DP rigorously and provably quantifies the **privacy loss** for published data, now and in the future
  - There is no more privacy loss - no matter what you do to the published DP data - as long as the original data are untouched
  - In particular, linking DP data to external data does not leak privacy
  - If original data are re-accessed (e.g., in a secure data enclave), privacy is leaked, and this must be reflected in the accounting

WG comments on Bureau's choice of DP

- DP is relatively new but is the current "gold standard" in industry and academia
- DP is an area of active research
  - many unanswered theoretical and methodological questions
  - many computational challenges
- Bureau is forward-looking in adopting DP in its Disclosure Avoidance System (DAS) to protect confidentiality
- *WG commends the Bureau for its serious commitment to modern and future-proof privacy protection and its development of Differential Privacy protocols*

DP requires a privacy-loss budget

- DP requires explicit choice of a **privacy-loss budget**, quantified with a parameter, ε (epsilon)
- ε is a parameter in a double-geometric noise distribution: high values imply more accuracy/less privacy, low values imply less accuracy/more privacy



- ε determines a budget in the sense that there is a total value that must be allocated across all products: spending more to gain accuracy for one product means less to spend on accuracy for other products
- Budget choice is a complex, consequential, irreversible, **mission-critical decision**

Budget is allocated across 2020 Census data products

- Possible state-level estimate of undocumented immigrants? (to accompany invariant state-level population counts)
- Group I Data Products
  - PL 94-171 Redistricting data file
  - Citizen Voting Age Population by Race and Ethnicity (CVAP)
    Demographic Profiles
  - Demographic and Housing Characteristics file
- Group II Data Products
  - Detailed race, ethnicity, and tribal data
  - Person-Household joins

Census process for implementing DP

- The Bureau could not apply off-the-shelf technology for implementing DP at "2020 Census scale"
- DP releases from 2020 Census must satisfy complex requirements:
    - **tabular summaries** that are nonnegative counts with internal consistency (tables add up) and no nonsensical results (e.g., need at least as many people as there are occupied households)
    - **invariants** to be released without noise: state populations for apportionment (also count of housing units by block and occupied GC by block*type; others?)
    - a synthetic, **DP microdata** file
- The Bureau's **Top Down Algorithm** largely solves these technical problems

Top Down Algorithm (TDA), simplified

- **Input the microdata:** At each geographic level (starting with the nation and working down to blocks), have a non-negative, integer-valued “histogram” that represents the microdata
- **Create tables:** Determine the set of queries to be protected at this level: the detailed tabular summaries to be published
- **Protect the tables:** Add noise to the queries, to get a noisy table
- **Create protected microdata:** Find the new microdata that come as close as possible to generating the noisy tables: a big, complicated optimization, with many constraints *[this is like a reconstruction attack]*
- **Repeat at the next level:** Output microdata now become input for the next, finer level of geography

Improvements to post-processing in TDA

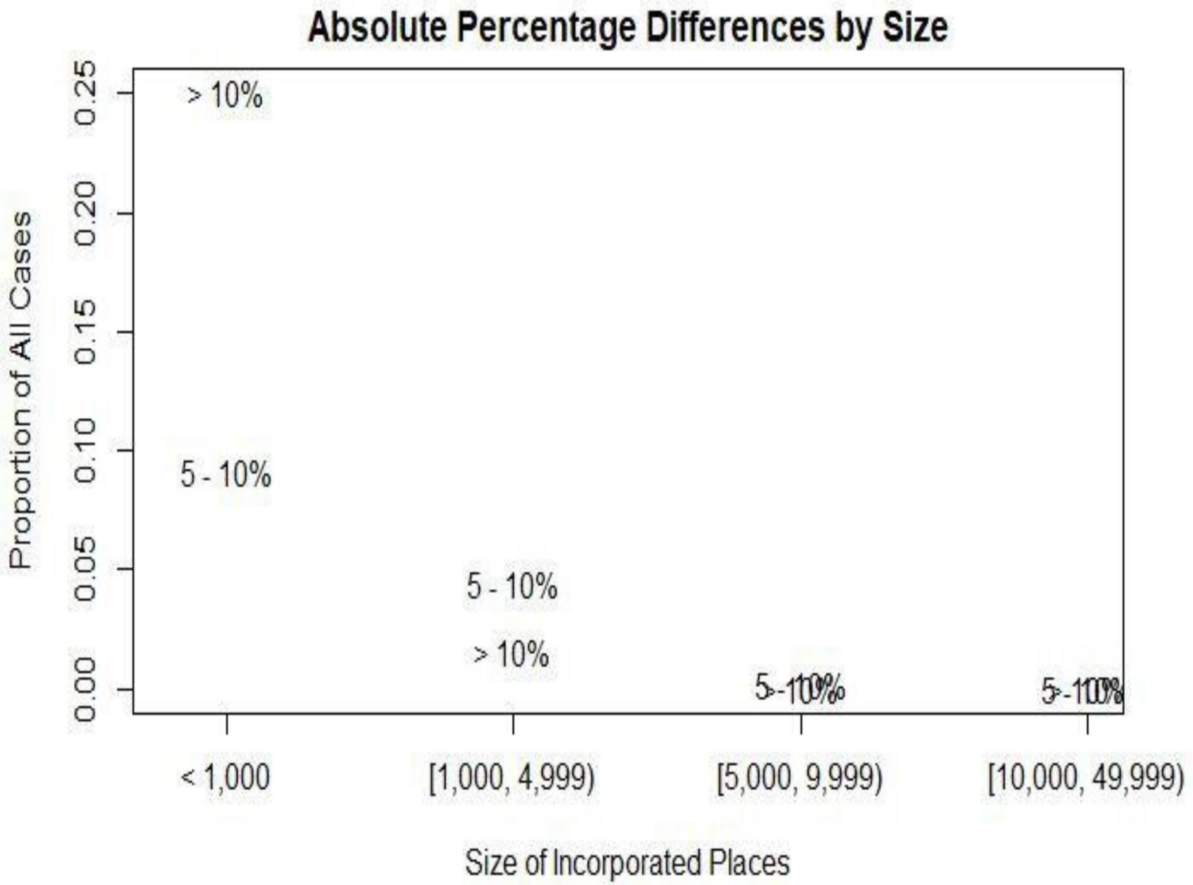
- 2010 Demonstration Data Products:
  - Protected tabulations based on 2010 Census
  - Released in October 2019 for evaluation by user community
  - Users noted serious distortions in DP data, via email and at December 2019 National Academy of Sciences Committee on National Statistics workshop
- Bureau's response was refinement of TDA post-processing
  - Now conducted in a series of passes within geographic levels
- TDA showed substantial improvements in various quantitative metrics after revised post-processing, based on results released in May
- Further refinements to the post-processing were expected to be shown in revised metrics, but no updates have been released as yet

TDA and post-processing, continued

- TDA with revised post-processing is now fully functional
  - Substantial effort to improve post-processing following initial release
  - Caution: fully functional is not necessarily optimal
- *WG notes that the Bureau’s implementation of DP at 2020 Census scale via TDA is a **major technical achievement***
- *WG found that the Bureau used an **exemplary development process** for its DP algorithms and code, following current best practices and making new contributions to the field*

Greatest impact on privacy/quality for smallest domains

- For a given privacy level, added noise has a relatively larger impact on smaller domains
- Because added noise can yield negative counts in small domains, rounding up results in **positive biases**
    - Relatively large positive biases because counts are small
    - Balanced by relatively small negative biases on larger domains



TDA and the privacy-loss budget

- TDA requires as input the overall privacy-loss budget (PLB) and its allocation across the geographic hierarchy (nation, states, counties, tract groups, tracts, block groups, blocks) and across all query sets
- PLB allocation affects privacy vs. accuracy for all census data products, with impacts on all uses of census data
    - Largest impacts on smallest domains
- PLB allocation is a complex, consequential, irreversible, mission-critical decision
    - To be determined by the Bureau's Data Stewardship Executive Policy Committee (DSEP)
    - **Process by which DSEP will make these decisions and what factors will be considered are not apparent to the WG**

Bureau’s 05/27/2020 timeline

- **5/27/2020** Disclosure Avoidance System (DAS) Accuracy Metrics from Sprint II (revision of 2010 Demonstration Data Products)
  - (Additional runs were planned every ~6 weeks, *but we have not actually had any releases since Sprint II*)
- **September 2020** DSEP sets final algorithm design and set of invariants
- **March 2021** DSEP sets final privacy-loss budget and its allocation
- **April 2021** DAS production run for Group I products
- **Summer 2021** Begin releasing Group I products
- **Winter 2022** Begin releasing Group II products

Bureau’s revised timeline?

August 26, 2020: “The Census Bureau is applying all of its resources to ensure we are able to meet the legal deadline for producing an apportionment count that is complete and accurate. After we solidify our plans to accomplish this goal, we will turn our attention to identifying the activities and schedules needed for producing and delivering accurate and complete Redistricting (P.L. 94-171) data to the states, the District of Columbia and Puerto Rico as expeditiously as we can. This includes finalizing the implementation of the new Disclosure Avoidance System (DAS). ”

- *WG commends the Bureau for maintaining a centralized location for updates: https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/2020-das-updates.html*
- *WG would appreciate any updates on the timeline for implementation of DP*

Task 1: Summary of Use Cases for 2020 Data Products

Goal: **Evaluate the Census Bureau's collection of use cases:**

- What are we missing?
- What haven't we considered?
- *Especially important for uses of Group II data products (detailed race, ethnicity, tribal data and person-household joins)*

Background: 2020 Census Data Products Planning Crosswalk
Summary of collected use cases (Census Bureau will email to members)

Deliverable: **Presentation on your review of use cases at Fall NAC/CSAC Meeting**

Sources for soliciting use cases, 2018-2019

- Federal Register Call (SF1 tables): July - September 2018
- Comments from Users of the 2010 Demonstration Data Files (October 2019)
- Committee on National Statistics Workshop on 2020 Census Data Products (Dec 2019)
- Comments from Disclosure Avoidance System Updates website, including metrics release.
- Comments from users of the new Demonstration Data Files, produced in partnership with Committee on National Statistics
- Ad hoc letters from stakeholders
    - National Council of State Legislators, Washington state
    - Federal-State Cooperative for Population Estimates members
- Exchange at panels at external events with Census staff participants
- CSAC/NAC workgroups

Evaluate the Bureau’s collection of use cases, I

- *WG applauds the Bureau’s efforts to seek input on use cases from multiple sources*
- *WG appreciates the Bureau sharing the collection of Federal Register use cases and encourages the spreadsheet to be posted publicly.*
- Many important use cases are represented
- Research community is fairly well represented
- What cases are not represented? How do you know?
    - Frame development and refinement? NRFU?

Evaluate the Bureau’s collection of use cases, II

- Overarching strategy for determining representativeness of use cases is unclear
- We still lack a sufficient picture of repercussions of differential privacy
    - Creating ripple effects from other Decennial Census-derived data sets
    - Altering distribution of federal grant programs
    - Determining context for government regulations
- Looking across all stakeholders (public, community, private sector,...)
    - Not clear that all are aware of these potential repercussions of DP
    - High barriers (time, complexity) to engagement
    - Many are overwhelmed in responding to pandemic - limited energy to engage

Key types of uses

- Political representation and redistricting
- Funding allocation & program eligibility- *e.g. USDA programs by rurality; Indian Housing Block Grant; states to cities/counties*
- Legal mandates- *e.g. Environmental Justice reviews*
- Regulatory practices- *e.g. Fair Lending Laws & Regulations*
- Planning- for programs & services - *e.g. schools, elder care, emergency management*
- Research
- Private Sector/Business

Analysis of Federal Register responses: What use cases might be missing?

- Of the 15 federal executive agencies, only eight appear in the use cases.
- Of the 13 agencies in the federal statistical system, only five explicitly appear among the use cases.
- Many states use Census numbers for sub-state funding allocations and other purposes, but only 15 states appear among the use cases (and Puerto Rico does not appear either).
- Of the thirty largest US cities by population, only three appear by name among the use cases (New York, San Diego, San Francisco).
- Private sector investment decisions completely absent

Methodical use-case catalog development

- Developing (and maintaining) a public catalog of all federal and state uses of Decennial Census data would be useful for many purposes, and would help in determining epsilon and the privacy-loss budget allocation, and would aid in consideration of DP impacts.
- ***DRAFT RECOMMENDATION (for full CSAC consideration): The Bureau should take substantially more time to catalog methodically the use cases of census data, including funding allocations, legal mandates and regulatory practices, across all agencies of the federal government as well as at state and local levels.***

Rigorous analysis for priority use cases

- Additional rigorous analysis needed for different use cases.
    - Analyses of impacts on funding formulas for federal agencies and congressional staffers
    - Analyses of impacts on legal mandates and regulatory practices
- ***DRAFT RECOMMENDATION (for full CSAC consideration): Once the Bureau has more thoroughly cataloged important use cases of census data, they should conduct analysis of the impact of Differential Privacy for priority use cases (funding, legal, and regulatory at all levels of government). An example of such analysis is Variability Assessment of Data Treated by the TopDown Algorithm for Redistricting (Wright and Irimata 2020) .***

Task 3: Metrics for Assessing the Impact of the 2020 DAS on Data Accuracy

Goal: **Evaluate the Census Bureau's Accuracy Metrics**

- Are we capturing the important use cases?
- Are the metrics effective for evaluating accuracy of the DAS runs?

Review metrics applied to new DAS runs and assess "fitness for use"

Background: **Accuracy Metrics**

- **Overview**
- **Baseline – 2010 Demonstration Data Products**
- **Sprint II Run**

Deliverable: **Public can submit comments to 2020DAS@census.gov**

Presentation on your assessment at Fall NAC/CSAC Meeting

Metrics for 2020 Census

- Metrics are essential for users to judge quality and fitness of use
- Releasing any data summary leaks some privacy
- Error metrics computed for 2020 data products are data summaries that leak privacy:
  - WG thinks error metrics computed from 2020 will not be released [to be verified]
- Instead, users have access to metrics for 2010 data products
- By analogy (?), expect similar behavior for DP 2020 data products as seen in DP 2010 demonstration products

2010 Metrics so far

- October 2019: 2010 Demonstration Data Products released for evaluation by user community
- May 2020: Sprint II - adjusted post-processing in comparison to prior version, in response to user critiques
  - No full demonstration product, but detailed summary metrics comparing Census 2010 data to Sprint II version of DP 2010 data
  - Commonly-used summary metrics at various geographies for key variables
  - ***Some variables (use cases) should be included - (e.g. housing vacancy status- seasonal homes)***
  - ***Some geographies (use cases) should be included/better represented (zip codes, county subdivisions/minor civil divisions- political units in Northeast and Midwest)***

Privacy-protected 2010 microdata

- July 2020: Bureau released privacy-protected 2010 microdata that allows users to construct whatever metrics they like or analyze for any use.
  - ~ 309 million observations (individual records)
  - Most informative, least accessible
  - User community hasn’t had time to process- technical issues - huge data demand and expertise required
- August 2020: IPUMS put together summary metrics based on 2010 microdata from Sprint II (May 2020)
- Additional versions (sprints) have not been released
  - Concern that needed refinements may be behind schedule

Metrics released with 2010 demonstration product

- **Mean Absolute Error (MAE):** arithmetic mean of absolute errors
- **Mean Numeric Error (ME):** arithmetic mean of errors
- **Root Mean Squared Error (RMSE)**
- **Mean Absolute Percent Error (MAPE):** arithmetic mean of absolute relative errors
- **Mean Percent Error (MALPE):** arithmetic mean of relative errors
- **90th Percentile of Absolute Percent Error**
- **Coefficient of Variation (CV):** RMSE / (mean of characteristic)
- **Total Absolute Error of Shares:** average absolute difference of proportions

Metrics limitations: Detecting bias

- We know of some systematic bias in error due to nature of TDA:
  - Smaller population areas overestimated while larger pop places are underestimated.
- Current metrics (outside of microdata) don't allow for evaluation of other biases that might undercount or overcount…
  - … specific groups (by race, age, ethnicity, etc.)
  - … within certain types of geographies (e.g. rural/urban, spatially proximate, race by age, etc.)
- The microdata are useful - in principle, can compute anything you might want - but the barriers to access are high, even for sophisticated users

Metrics limitations: Assessing outliers

- Outliers need to be protected, so metrics offer little to help identify them
- Can anything be said about patterns in the cases where error is really high?
- Is there systematic bias in number/size of outliers for certain types of places/domains?
- Current error metrics don’t leave much room for evaluating differential outlier behavior

Suggestion: Using 2010 metrics for 2020 analysis

- Generalized variance functions (GVFs), an old Bureau standby:
  - Compute point estimates and standard errors for many responses in a survey
  - Build a regression model to predict standard error as function of point estimate and sample size
  - Publish these GVFs instead of standard errors
  - Standard error is then approximately GVF(point estimate, sample size)
- Generalized metrics functions (GMFs)?
  - Compute metrics for many 2010 demonstration products
  - Build a 2010 regression model to predict metric as function of DP point estimate, cell size
  - Publish these GMFs for 2020 instead of actual 2020 metrics
  - 2020 Metric is then approximately GMF(2020 DP point estimate, cell size)

Suggestion: Split up MALPE

- MAPE is
  - {(sum of positive relative errors) - (sum of negative relative errors)} / N
  - Direction is lost due to absolute value
- MALPE is
  - {(sum of positive relative errors) + (sum of negative relative errors)} / N
  - Errors can cancel, so magnitude of errors in each direction are lost
- More informative to report separately:
  - (mean of positive relative errors), (mean of negative relative errors)
  - (number of positive relative errors), (number of negative relative errors)
  - Can reconstruct MAPE and MALPE from these values

Suggestion: Consider alternatives to TAES

- **Total Absolute Error of Shares (TAES):** For some geographic partition (g=1,2,...,G), TAES is the sum of absolute differences between original proportions Q(g) and privacy-protected proportions P(g):
  - $\Sigma|Q(g) - P(g)|$
  - “goal is to provide a measure of the distributional error”
- But there are many established methods to compare two discrete probability distributions; e.g.,
  - **Kullback-Leibler divergence**: $\sum Q(g)\ln\{Q(g)/P(g)\}$
  - Potential links to justification from probability and information theory

Examples of remaining concerns in Sprint II, (1)

- General improvement in Sprint II over 2010 Demonstration Data
- Remaining concerns: there are still large errors for some commonly-used geographies and variables
- **These may have implications for funding allocation and planning**
- ***Ex. 1: Incorporated places- Total Populations***
  - Funding distributed to incorporated places by population size
  - MAPE (Mean Absolute Percent Error) = 9%; MAE (Mean Absolute Error) = 55
  - ~ 2,700 places with % error 5-10%
  - ~ 4,700 places with % error >10%
- ***Ex. 2: Total Populations by Age***
  - Age 65 plus: County: MAPE= 7%; MAE= 231; Incorporated Place: MAPE= 30% ; MAE= 89
  - Under Age 5: County: MAPE= 8%; MAE= 84; Incorporated Place: MAPE= 46% ; MAE= 40

Examples of remaining concerns in Sprint II, (2)

Ex. 3: Blocks- Total Populations

- Blocks are used to build bigger, non-standard geographies (e.g., for redistricting)
- Problem with systematic bias. Current metrics don't offer ability to test if bias is systematic or how grouping spatially adjacent blocks impacts error
  - Example A: State of Washington: PPMF has 15,253 people in blocks that had zero population in SF1 (from PPMD analysis)
  - Example B: Total Population in blocks classified as urban vs rural. Impacts calculation of % rural/urban at county and state level, implications for funding (USDA, etc.)
    - Urban blocks: MAPE=51%; MAE= 8; MALPE= 30%
    - Rural blocks: MAPE= 78%; MAE= 4; MALPE= 55%

Potential for accumulation of bias

- Small domains are more likely to suffer relatively large, positive bias due to non-negativity constraint and post-processing adjustments
- Adding up small domains (e.g., blocks) to create new (off-spine?) geographies may accumulate bias and this may be consequential
  - Total of B blocks grows with order $O(B)$; Standard deviation of error sum total grows with order $O(\sqrt{B})$ and is inconsequential; Bias of error sum total could grow with order $O(B)$
- One way to assess this bias is to compare to privacy-protected data **before post-processing** (which might even be negative)
- These data are unbiased for the private summaries and have known error distributions

Useful data for bias assessment

- A useful tool for bias assessment is comparison of privacy-protected data before post-processing to data after post-processing
- ***DRAFT RECOMMENDATION (for full CSAC consideration): To facilitate assessment of bias properties for the privacy-protected data, the Bureau should release the non-post-processed measurements used in TDA, which are unbiased estimates with known error distributions.***

Missing metrics?

- **Existing metrics** are not reported for some important, missing use cases
  - Ex. by geography- Minor Civil Divisions, Zip Codes
  - Ex. by variable- Housing Vacancy (seasonal housing)
  - Ex. by geography & variable- race/ethnicity by block (used for environmental justice, USFS)
- Some **new metrics** might be suggested when considering other missing use cases
- ***DRAFT RECOMMENDATION (for full CSAC consideration): The recommended use-case catalog development and rigorous analysis for priority use cases may suggest the need for new metrics. The Bureau should revisit the list of metrics periodically as the use-case catalog and analyses evolve.***

2020 metrics from 2020 data?

- Will there be any measures of error published for 2020 Census from 2020 Census data? (at some cost to privacy-loss budget)
  - It is not clear that users are aware that metrics leak privacy, and may not be released
- Whether or not any metrics are published based on 2020, users will need to do some extrapolation from 2010 metrics to 2020 uses
- ***DRAFT RECOMMENDATION (for full CSAC consideration): The Bureau should make clear what, if any, metrics for 2020 will be computed from 2020 data. The Bureau should make readily available tools (like the suggested GMFs) for extrapolating from 2010 demonstration metrics to 2020 use cases.***

Additional metrics

- *The WG appreciates the Bureau's efforts in not only releasing a suite of metrics, but also releasing microdata*
- **Too much:** microdata are massive and complex, and the barriers to using these data are high
- **Not enough:** The set of metrics so far is reasonable, but more information is needed to adequately assess outliers, remaining biases, etc.
- **Just right?**
  - Consider releasing the existing metrics for more domains
  - Consider releasing more distributional information: overall means can hide information (as in the MALPE example), but means within bins (e.g., defined by quintiles) would reveal more information

Task 2: Prioritize Use Cases

for Allocation of Privacy-Loss Budget and Adjustment of the Statistical Processing Algorithms

Goal:

For Group I data products (PL 94-171, Demographic Profiles, and Demographic and Housing Characteristics file)

- Recommendations on how PLB should be allocated across data products and tabulations

For Group II data products (detailed race, ethnicity, and tribal data and person-household joins)

- Recommendations on necessary geographic and variable detail for Group II data products

Background:

Disclosure Avoidance System (DAS) updates website
2020 Census Data Products Planning Crosswalk

Deliverable: **Presentation on your recommendations at Fall NAC/CSAC Meeting**

Privacy-loss budget allocation in the TDA

- Current implementation of TDA has homogeneous allocation:
  - epsilon shares are constant across query sets within geographic levels
- Geography in the TDA follows the geographic spine:
  - Nation, states, counties, tract groups, tracts, block groups, blocks
  - Invariant population counts at state level
  - Invariant household and occupied group quarters counts at block level
- Geography not in the TDA:
  - AIAN areas, minor civil divisions, incorporated places, ZIP codes, etc.
- Query sets within levels:
  - total population, relgq
  - votingage * hispanic * cenrace: 252 queries
  - age * sex * hispanic * cenrace: 29, 232 queries

Prioritization of use cases in the PLB allocation?

- WG asked to advise on prioritization of use cases in the allocation of the privacy-loss budget (PLB) across data products
- WG does not know the implications of PLB allocation for **privacy**
  - How does the allocation across geographic levels affect re-identification risk?
  - How does the allocation across query sets within levels affect re-identification risk?
  - How does likely differential nonresponse affect re-identification risk?
  - What level of re-identification risk is considered “acceptable”? Or at least, does the Bureau have an answer to this question?

PLB allocation, (2)

- WG asked to advise on prioritization of use cases in the allocation of the privacy-loss budget (PLB) across data products
- WG does not know the implications of PLB allocation for **accuracy**
  - Constraints and invariants in TDA greatly complicate any accuracy assessments
  - No good way to assess theoretically; assess empirically across allocations?
  - WG does not know the full suite of use cases, their accuracy requirements, their differential bias
  - Each use case has its own accuracy requirements to determine fitness-for-use
  - Each priority use case requires rigorous analysis
  - Accuracy requirements will sometimes be in conflict across use cases

PLB allocation, (3)

- We were asked to advise on prioritization of use cases in the allocation of the privacy-loss budget (PLB) across data products
- WG does not know the implications of PLB allocation for **privacy-accuracy tradeoff**
  - Overall required privacy and privacy-accuracy tradeoff are mission-critical decisions
  - WG does not know how the Bureau will make these decisions, what factors are being considered as the Bureau makes these decisions, and when these decisions will be made
  - Given a choice of privacy/accuracy balance, CSAC might have suggestions on allocation

Thoughts on priority for use cases

- Some use cases rely on invariants and do not use up privacy budget
- Some use cases rely on large-domain estimates and any reasonable PLB allocation should have little effect on accuracy
- Some use cases could evolve in response to added noise; e.g., finding supplemental data sources or adjusting formulae (soft vs. hard threshold)
- Some small-domain use cases will be more robust to DP perturbations than others
- All priority use cases require empirical analysis with objective measurement and evaluation using an array of fitness-for-use measures
- Any proposed PLB allocation across use cases must be evaluated for re-identification risk

Priority for use cases

- Based on WG examination of collected use cases and our assessment of potentially missing use cases, we propose the following priority:
- ***DRAFT RECOMMENDATION (for full CSAC consideration). Overall, the privacy-loss budget should be prioritized toward the most important use cases in this order:***
  - ***Government funding (federal, state, local)***
  - ***Legal mandates and regulations***
  - ***Community planning (children's & elder services, infrastructure)***
  - ***Academic research***
  - ***Citizenship***

Citizenship data?

- Bureau planned to link undocumented individuals from administrative records to 2020 Census data prior to privacy protection, for December release.
- Bureau is developing estimates of the number of citizens in each block based on administrative records for CVAP for release in 2021.
- This means citizenship status would receive a share of privacy-loss budget
- ***DRAFT RECOMMENDATION (for full CSAC consideration): If any citizenship variables are part of the December release or CVAP release, the Bureau should assign to these variables a very small part of the privacy budget: no more than …***

Inhomogeneous allocation of PLB?

- In current implementation of TDA, epsilon shares are constant across query sets within geographic levels
    - E.g., (county-level) (votingage * hispanic * cenrace) gets (12% for county) (29% for query set) (overall ε), for every county
    - In principle, epsilon could be allocated in many other ways
- Geographic levels need not receive homogeneous allocation, even if algorithmically convenient
    - E.g., group by population size; rural/urban or regional differences in important variables
- Query sets need not receive homogeneous allocation
    - Could race/ethnic categories be collapsed so as to allocate more privacy loss budget to more commonly used groupings and reserving little for more specific groupings? This might be especially useful at small geographies.
    - OMB has six standard classifications: used in Pop Estimates program, ACS, etc.

Thoughts on PLB allocation and timeline

- Bureau's DP implementation is operating on an ambitious timeline under any circumstances
  - Census 2020 is in the field during interesting times
- Bureau is operating under enormous time pressure to make the incredibly consequential and irreversible decision on privacy-loss budget allocation
- Many implications of the decision for privacy, accuracy, and fitness-for-use are currently unknown
- There are **known use cases** for which error has improved but is still high
  - What is good enough for known use cases?
  - Need rigorous analysis to assess accuracy and fitness for use
- There are likely to be important **unknown use cases** and unheard users

Thoughts on PLB allocation and timeline, continued

- Currently incomplete understanding of privacy/accuracy tradeoffs
- Difficult to see how to make PLB allocation decisions given these unknowns
- Process by which PLB allocation decision will be made is unclear
    - How will the Bureau make this decision, and what factors are being considered?
- Whatever the choice of PLB allocation…
    - Need to estimate the re-identification risk to ensure sufficient privacy
    - Need to give users some way to assess fitness-for-use
    - Need to have a backup plan (e.g., allocate some privacy budget) for the future, in case DP data are not fit for some important use cases

Recommendation to take more time

- Previous draft recommendations suggested (a) Methodical use-case catalog development and (b) Further rigorous analysis for priority use cases
- ***DRAFT RECOMMENDATION (for full CSAC consideration): The recommended use-case catalog development and rigorous analysis for priority use cases are important for informing how to allocate the privacy-loss budget across uses. The Bureau should put off additional releases after the December apportionment release to allow time for these analyses.***

Discussion